Decision-Making and Learning - Comparing Orthogonal Methods to Majority-Voting

Daniel W. Repperger$^{\nabla}$

$^{\nabla}$ Air Force Research Laboratory, AFRL/HECP, WPAFB, Ohio 45433, USA, D.Repperger@IEEE.ORG

Abstract. A study on learning and decision-making methods was conducted by comparing an orthogonal methodology of manipulating data versus that of a majority-voting procedure. The latter method has recently become popular in the literature involving applications such as pattern recognition. To evaluate the differences between the proposed methods, data from a multidimensional paradigm involving decision-making and learning are analyzed. A number of basic concepts from estimation and information theory are first discussed to understand both the motivation and the underlining issues involved in conducting this study.

Key Words: Decision Making Methods

I. INTRODUCTION

Learning and decision-making are processes that adapt and are highly multidimensional [1]. Also when developing autonomous systems, there is considerable interest in adaptability as an intelligent means of modifying behavior as new data are acquired. Much like learning, decision-making to improve the quality of information has similar and related issues to designing intelligence in autonomous systems [1,2,3,4]. In a recent study [5], it has been demonstrated that it is possible to build a decision-making scheme from a "bottoms up" approach starting with a vector of orthogonal classifiers. Alternatively, a different approach involving classification and learning procedures occurs in pattern recognition schemes [6] where a scalar measure (majority-voting) can be compared to the hyperplane method as discussed in [5]. This paper will cover the basics of a decision-making process and how it can be generalized to learning by extrapolation of the techniques presented here. Both methods are highly adaptable, which is of interest in a number of special applications, and, in particular, for intelligent control methods involving the design of autonomy. First it is important to discuss some well-known results from estimation and information theory which motivate the orthogonal approach discussed here.

In estimation theory (e.g. in Kalman filtering) the concept of orthogonal projection is well-known. An optimal estimator is recognized as having its error vector orthogonal to the direction of the measurement signal. Another interpretation of this result is that the residuals (difference between the data and the estimator) should contain zero information (the residuals are random) and are not correlated with the state estimate [7]. Hence one can view learning as a process of making the residuals white (containing no information) and the error of a state vector remaining orthogonal to the measurement set. Thus learning can proceed, as new data are received, by updating the estimator, accordingly, so that the resulting residuals still contain minimal information. This is also consistent with information theory concepts in which the greatest information is contained in the most unlikely event and there is little new information in an expected event [8].

When multiple channels of data tell the observer their potential classification of a particular object, the decision can be predicated on the orthogonal approach or possibly on the majority vote of scalar classifiers. There are two distinct points of view:

(1) The first and traditional method (vector) is that an optimal estimator can be built which employs an orthogonal method described above. As new data arrive, the estimator is adapted so that the resulting error vector remains orthogonal to the measurement set. This methodology is not necessarily a scalar process and hyperplanes can describe the estimator when any number (n) of channels of data are available.

(2) The second possibility (scalar) is that a majority-voting scheme could be employed. This differs from the method (1) because of n (initially assumed to be odd) channels of data could each individually select (binary decision rule) their choice of a decision on the classification of an object. The overall decision is then based on the majority of the decisions. This second method is a scalar mapping; the first method involves a hyperplane or vector methodology. It has been shown mathematically [6] that the second method can be as effective or better than the first method in certain situations. This paper will examine the relevant details why learning or decision-making may benefit from a majority viewpoint in contrast to an orthogonal perspective. First the basics of each of these processes are reviewed.

II. Examples Considered

To better understand the relevant issues, the basics are reviewed utilizing well-known results involving information theory, Kalman Filtering, and orthogonal pattern recognition procedures. The goal is to compare both across and within different methodologies to see similarities and differences on why certain methods may help adapt in learning and why a majority-voting scheme has some merit. The first example arises from the basic mathematical discussion of orthogonal projection.

2.1 Optimality and Orthogonal Projection

To provide the background to this approach, it is first instructive to show the fundamental relationship between optimality and orthogonal projection. Given a linear space X with inner product <x, y> defined for any two elements using the L_2 norm:

$$\| x \| = < x,x >^{1/2} \qquad (1)$$

A fundamental theorem is borrowed from the classical literature in this area [9].

Theorem 1: $\| x - \hat{y} \|$ is a minimum for all $y \,\varepsilon\, M$ (the measurement set) , i.e.

$$\| x - y \| \geq \| x - \hat{y} \| \quad \forall\ y \,\varepsilon\, M \qquad (2)$$

if and only if $(x - \hat{y})$ is orthogonal to all $y \,\varepsilon\, M$, i.e.:

$$< x - \hat{y} \ , y > = 0 \quad \forall\ y \,\varepsilon\, M \qquad (3)$$

Proof:

First assume equation (3) is valid, then for any $y \,\varepsilon\, M$,

$$\| x - y \|^2 = \| (x - \hat{y}) + (\hat{y} - y) \|^2 \qquad (4)$$

$$= \| x - \hat{y} \|^2 + 2 < (x - \hat{y}),\ \hat{y} - y > + \| \hat{y} - y \|^2 \qquad (5)$$

where each $(y - \hat{y}) \,\varepsilon\, M$. But from equation (3), the middle term of (5) vanishes yielding:

$$\| x - y \|^2 = \|(x - \hat{y})\|^2 + \| (\hat{y} - y) \|^2 \qquad (6)$$

$$\geq \quad \| (x - \hat{y}) \|^2 \qquad (7)$$

with equality if and only if $y = \hat{y}$. To complete the proof, (assume (3) is not valid) and that $\hat{y}$ minimizes $\| x - y \|^2$ for all $y \,\varepsilon\, M$, hence there exists some $y_1 \,\varepsilon\, M$ such that:

$$< x - \hat{y} , y_1 > = \alpha \neq 0 \qquad (8)$$

Then: $\| x - \hat{y} - b y_1 \|^2 =$

$$\| (x - \hat{y}) \|^2 - 2\ a\ b + b^2 \| y_1 \|^2 \qquad (9)$$

Thus it appears that by appropriate choice of b it is possible to make the combined total of the last two terms of (9) negative, thus contradicting the minimality of $\hat{y}$. Hence such an element y_1 of M cannot exist and this shows the optimality criterion.

Remark:

The relationship between optimality and orthogonality is immediately evident. The orthogonal component **y** clearly minimizes the function:

$$J_1 = \min \| \mathbf{x} - \mathbf{z} \| \qquad (10)$$

over the set of vectors **z** in *M* as illustrated in the proof of this theorem. Thus if the goal is optimality (in the sense of minimum distance), then the orthogonal projection provides a viable solution. Next, this concept is described in terms of the well-known Kalman filter and the principle of orthogonal projection.

2.2 An Example from Estimation Theory (Kalman Filter):

The well-known Kalman filter was derived using the concept of orthogonal projection [7,9,10]. For brevity, only the basic details are presented here. Let $\hat{x}$ denote the estimate of the state vector *x* as the solution of the optimal linear filtering problem. The error is $\tilde{x} = x - \hat{x}$. Using the expectation operator notation, the optimal estimator at time t_1, provided by measurements *z(t)* up to time *t*, satisfies the following two important properties:

(a) $E\{ \hat{x}(t_1 \mid t) \} = E\{ x(t_1) \}$

(b) $\min E\{ \| \tilde{x}(t_1|t) \|^2 \}_B$ is achieved.

The matrix B is a positive definite matrix.

The orthogonal projection lemma relates to the above conditions as follows:

Orthogonal Projection Lemma for the Optimal Linear Estimator

The optimal estimator satisfying conditions (a,b) above also satisfies the following orthogonality condition [7,9,10]:

$$E\{ (\tilde{x}(t_1|t))(z(t_1)) \} = 0 \qquad (11)$$

Remark: The Optimal Linear Estimator can also be derived from Theorem 2 [10]:

Theorem 2:

A necessary and sufficient condition for the linear estimator $\hat{x}$ to be the least squares (minimum variance) estimate is that

$$E\{ \hat{x}(t_1|t) \} = E\{ x(t_1) \} \qquad (12)$$

$$E\{ (\tilde{x}(t_1|t))(z(t_1)) \} = 0 \qquad (13)$$

In other words, if the estimator is unbiased (12) and orthogonal (13) to the measurement set, this is sufficient to minimize the least squares deviations. Hence orthogonality, linearity, and being unbiased are sufficient to guarantee optimality. We represent this concept in Figure 1 which portrays the error signal ($\tilde{x}(t_1|t)$), the measurement vector $z(t_1)$, and their orthogonal relationship. There is an interesting geometric interpretation in Figure 1 which elucidates the concept considered in this paper.

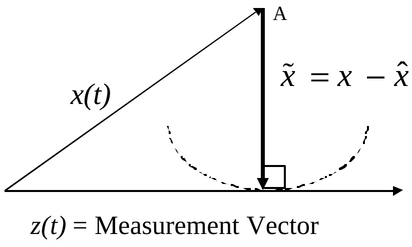

Figure 1 - Orthogonality Relationship between *z(t)* and $\tilde{x}$

Geometric Interpretation of Figure 1:

In Figure 1, one can view optimality in terms of a distance measure. Starting at point A as a center, a radius is drawn with length $\tilde{x}(t_1|t)$ as indicated by the arc. It has been known since the time of early Greece that the shortest distance from point A to the measurement vector *z(t)* (line) occurs if the radius is perpendicular to *z(t)*. Hence from a geometric perspective, the orthogonal projection is the minimum distance from a point to a line and the relationship between optimality and orthogonality is easily understood.

The next example is gleaned from information theory and insight is gained on how to relate this prior work on estimation theory to the information theory methods.

2.3 An Example from Information Theory

The approach here will be to synthesize a very complete model of an information channel to account for an assortment of possible losses and gains of information through a variety of processes [11]. The definition of the information *I(x ; y)* given by an observed event *y* about a hypothesis *x* can be specified in a probability sense as follows:

$$I(x\,;\,y) \;=\; \log_2 \frac{p(x|\,y)}{p(x)} \;\text{(bits)} \qquad (14)$$

The input set of *x*'s is defined as the discrete and finite set *X*, and the output set of *y*'s, correspondingly, is defined as *Y*. In figure 2, a flow graph (the information channel is inside the dashed box) is constructed with the following variables defined, accordingly:

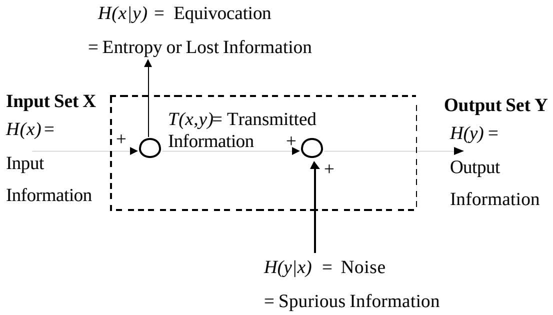

Figure 2 - The Flow of Information Through A Channel

H(x) = Input information in the set *X* (the information content of the set *X*).
H(y) = Output information in the set *Y* (the information content of the set *Y*).
H(y|x) = The noise added to the information channel (spurious information).
H(x|y) = The equivocation (entropy) which is the information about the input set *X* that might have been transmitted but was not.
T(x,y) = The transmitted information.

Some other interpretations of these key quantities can be stated. For example, *H(x)* is the input information provided in the source and *H(y)* is the output information received. The equivocation can be viewed as the average information still needed to specify an *x* exactly after the evidence *y* has been taken into account. The term average or expected value of information is derived from the fundamental definition of *H(z)* which is in the form of an expected value operation on information specified via:

$$H(z) \;\square\; \sum_i p(z_i)\,\log_2 \frac{1}{p(z_i)} \;\text{(bits)} \qquad (15)$$

Figure 2 displays the following equation representations of these different types of information measures:

$$H(x) - H(x\,|\,y) = T(x,y) = H(y) - H(y\,|\,x) \quad (16)$$

From figure 2, for a given information channel, the input information *H(x)* and the spurious information *H(y|x)* are generally fixed and specified. The best the designer can hope to accomplish is to reduce the uncertainty (*H(x|y)* = entropy or equivocation) by the choice of some design parameter or procedure. Two productive results occur if *H(x|y)* is reduced:

(a) The transmitted information *T(x,y)* is increased.
(b) The received or output information *H(y)* increases.

Hence reducing entropy or uncertainty, by any means possible, can only help to improve the quality of the decision-making or learning. For an autonomous or intelligent system, this can surely expand one dimension of intelligence by the means in which a decision is made. It will be shown in the sequel that the orthogonal procedure can also be viewed as an entropy reduction procedure.

To illustrate how decision-making can be realized from only an orthogonal approach, an example from pattern recognition is now introduced. Two approaches will be utilized to solve this problem. The first approach will be the construction of an orthogonal, hyperplane methodology. The second line of attack will introduce the procedure termed "majority-voting".

2.4 An Example from Pattern Recognition (Orthogonal Method)

A system is described which provides a means for improving the quality of information derived from a decision-making process by weighing certain multiple and alternative information channels. The method is applied to data estimating the cognitive workload state of a human operator dealing with a complex task using noninvasive sources of physiological data as a basis.

In recent years, as the proliferation of data becomes more and more persuasive, the challenge increases in designing systems that can process information in an innovative and efficient manner. The first system discussed in this paper has as a goal the improvement of the quality of information for making a decision from alternative (and multiple) sources of data. The potential data sources are first rank ordered in terms of their efficacy for making a binary decision. The next step is to combine two alternative data sources in a productive manner so as to glean out the highest quality information. By induction, the process then generalizes to multiple, alternative, data sources with the end goal of continuing to improve the decision-making process through the intelligent use of data. To illustrate the applicability of the approach, data relevant to the estimation of the state of an operator (human controlling an automated system) through the selection of certain, key, physiological signals provides a platform to test the efficacy of such a methodology [12].

As humans deal with highly automated and complex systems, it is sometimes desired to obtain estimates of elevated demands of cognitive workload as manifested by physiological signals that may be gleaned in a noninvasive manner. Once an identification of the operator in a high workload state is verified, the automation level of the system may be adjusted to maintain effectiveness of the mission [2,11]. Figure 3 illustrates the operator in a human-machine interaction system with physiological data being monitored. Figure 4 depicts the basis of the decision rule (low or high workload state) that will be investigated in this study with the goal of improving decision-making by using multiple channels of data in a productive sense. In Figure 4, the data displayed may be from as many as 43 possible physiological signals, which are obtained in a noninvasive manner.

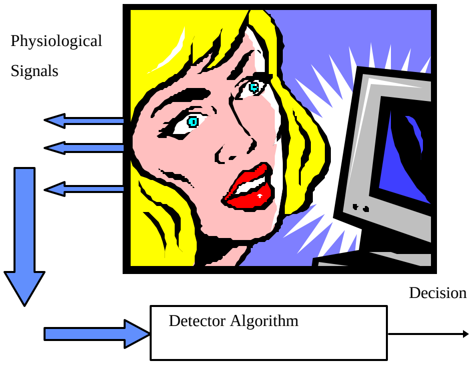

Figure 3- Physiological Signals to Detect Workload

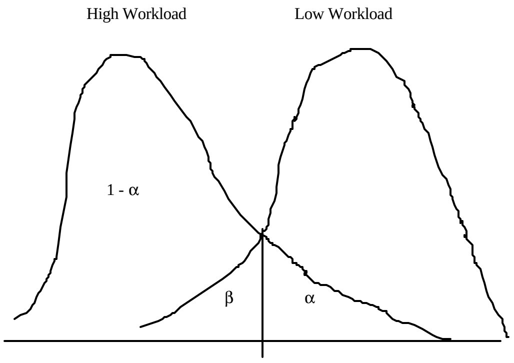

Figure 4- The Basis for The Decision Rule

2.5 The Statistical Decision Rule

Figure 5 portrays the ROC (relative operating characteristic) curve for data representative of figures 4 and 6.

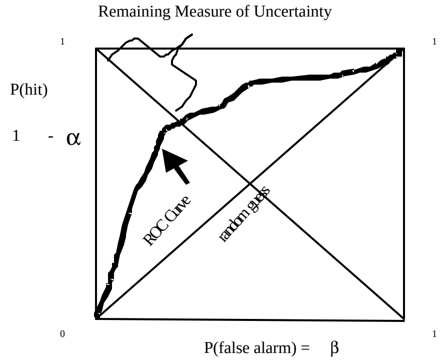

Figure 5 - The ROC Curve

The ROC was originally derived in signal detection theory, but has found widespread use in other areas. The plot in Figure 5 has

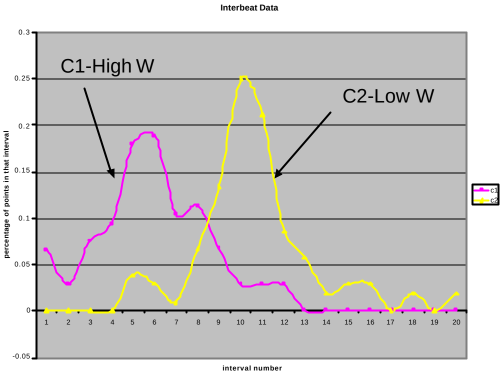

Figure 6 – Interbeat Heart Rate Data

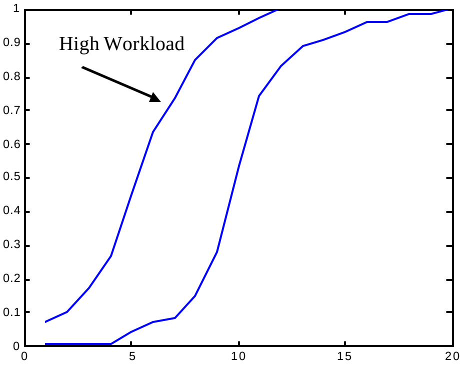

Figure 7 – CDFs for Figure 6 Data

as the dependent variable the term 1-α versus the independent variable β as derived from Figure 4. This may be viewed as the plot of the probability of a hit versus the probability of a false alarm in a binary decision rule [2,11,13] and can be shown to be the depiction of the two cumulative distribution functions of the densities of Figure 4. In an ideal decision-making process, the ROC curves moves upward to the left most diagonal (a measure of uncertainty, cf. Figure 5). Performance measures of such systems may be the minimum diagonal distance proximal to the upper left diagonal or the area under the ROC curve. An application to test the algorithm presented here is next described.

2.6 Testing the State of the Human Operator

From [12] there exist 43 possible data channels including physiological variables such as interbreath, interheart beat, and various electrode signals obtained as an operator performs a difficult task. Figure 6 illustrates the interbeat data for the two-workload conditions (high and low) and Figure 7 is the resulting cumulative distribution functions. Figure 8 is the corresponding ROC curve. Since the ROC curve is above the diagonal (random guess), this data variable is useful for predicting the state of the operator. The challenging problem discussed here is how to use two or more alternative data channels to improve upon the decision-making capability. After this procedure is illustrated for two channels, by induction, the process then generalizes to n channels.

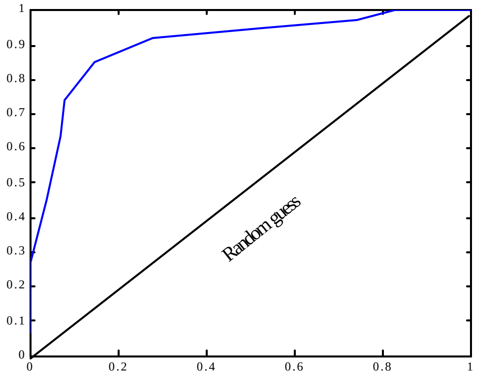

Figure 8 – Resulting ROC Curve for Figure 6 Data

2.7 The Orthogonal Algorithm

The algorithm to develop the decision rule has two steps:

Step 1: Rank order all data variables using the ROC curve.

Step 2: Select two or more data variables that yield a productive ROC curve, and then develop cross plots of the distributions. The decision rule is the hyperplane that separates the two distributions in an appropriate manner. Appropriate is based on an orthogonal projection between the centroids of the candidate distributions [14].

2.8 Implementation

Step 1 was implemented by plotting 43 ROC curves for all the data variables of interest. The efficacy (objective metric) was the minimum distance along the diagonal from the upper left corner to the ROC curve (cf. Figure 5). Thus all 43 data channels could be rank ordered, according to their ability to improve on the binary decision rule.

Step 2 was implemented by developing cross plots of two candidate distributions. The centroids were then calculated for each distribution. A line was drawn between the centroids. A perpendicular line was then constructed to separate the two distributions at a point determined by a ratio involving the distance of the respective ROC curves from their upper left corner on the diagonal in Figure 5. This decision rule then generalizes to a hyperplane as more variables are included. The overall decision rule (cf. Figures 9 and 10, for example) is that the selection is made of the high workload condition if the points fall below the hyperplane. Above the hyperplane is considered the low workload condition. The results then generalize to multiple channels of data and the decision rule is a vector based on ROC curves and hyperplane surfaces as shown in Figure 10 for any number of data channels. Also this method can be viewed as a means of reducing entropy by expanding the dimension set. In multiple dimensions, the entropy (lost information) is constantly reduced when the hyperplane includes more discriminate points in an n dimensional space.

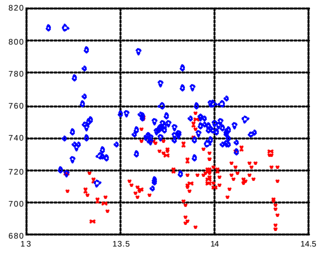

Figure 9 – Separating The Workload Data

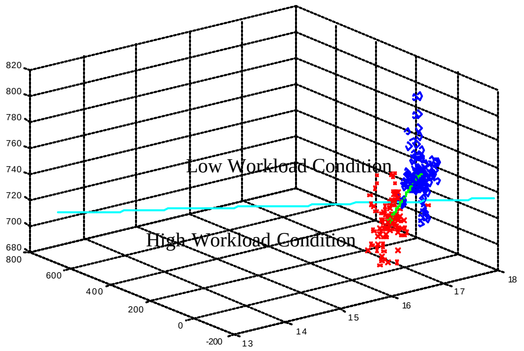

Figure 10 – Construction of A Decision Hyperplane

2.9 An Example from Pattern Recognition (Majority-Voting Procedure)

It has been shown mathematically [6] that a highly simple (scalar) algorithm can perform as well or better than an orthogonal scheme just described. Figure 11 displays a bank of classifiers (n is assumed to be an odd number). Each classifier makes an individual decision on the binary decision rule. The overall decision is simply the majority vote of these n classifiers. The advantages and disadvantages of this procedure are briefly described:

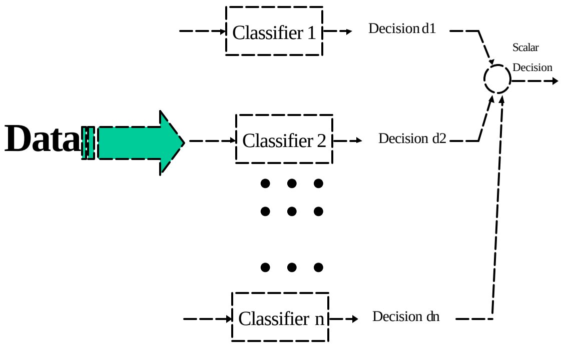

Figure 11 – Majority-Voting – A Scalar Decision-Making Process

2.10 Advantages of the Majority-Voting Procedure

Obviously, simplicity and the scalar nature of the process described in figure 11 is attractive, since computationally this process is much easier. Simplicity usually includes the attributes of reliability and robustness.

2.11 Disadvantages of the Majority-Voting Procedure

The disadvantage of the configuration in figure 11 occurs if the number of classifiers is small or does not fully represent the probability space concerning the important variables required in making a decision. If the number of classifiers $n \rightarrow \infty$, then it is obvious that the appropriate variables will be considered. This is analogous to the problem of “persistence excitation” in adaptive control theory. If, however, the system does not fully exploit the entire information set, then erroneous results may occur. Hence incorrect outcomes will occur if n is sufficiently small or does not include relevant information for making a key decision. We study the results with the application discussed previously.

III. Application to Experimental Data

Using data from [12] workload estimation of the human operator, the orthogonal method will be compared to a majority-voting scheme.

3.1 Comparison of the Orthogonal Approach to Majority-Voting

The comparison between these two sets of classifiers was conducted by studying three classifiers with a different data set as input to each classifier. This system was tested in an orthogonal sense as well as with the majority-voting scheme. The three selected physiological data sets from the 43 possible included: (1) interbeat (heart rate data), (2) electrode zero- alpha (the alpha brain wave from an electrode denoted as zero), and (3) electrode one- delta (the delta brain wave from the electrode denoted as number 1). It is noted that there were three nonelectrode data channels (interbeat, interbreath, and eyeblink) and 8 electrodes with 5 channels each of brain-wave data recorded. This gave a total of 43 channels of data possible to detect whether the operator was in a state of high or low workload. As these data were collected, the operator performed tasks, which were known to elicit a state of high or low workload by the task’s relative complexity and subjective comments collected.

The ROC curves of figure 5 were determined for all three data sets. The variable σ will be used to measure the distance from the diagonal to the upper left hand corner of the ROC curve along the vertical axis. Note $0.5 \geq \sigma \geq 0$ because a random guess line is described by the diagonal that goes from the (0,0) point to the (1, 1) in figure 5 and the efficacy of the estimator is the proximity of the ROC curve intersecting the diagonal going from (0,1) to (1,0). Four tests were performed. The classifiers were rank ordered by their σ values (the smaller σ is a better estimator). The orthogonal method and the majority voting method were both utilized to classify 210 points (106 in the high workload case and 104 in the low workload case). Table 1 shows the efficacy of the classifiers, alone. It lists the data utilized and the σ value for each classifier.

Table1–Efficacy of A Classifier Acting Alone

Classifier Number	Data Variable Utilized	σ from the ROC Curve
Classifier - 1	Interbeat (heartrate) data	0.15
Classifier - 2	Electrode 1- delta wave	0.27
Classifier - 3	Electrode – 0 – alpha wave	0.32

Thus as the classifier number increases, its ability to perform accurate decision-making degrades accordingly. The performance of these classifiers is now evaluated in both an orthogonal sense as well as in a majority-voting scheme. In Table 2, the errors e_1 represent the data points that were high workload but were wrongly classified as low workload. The errors e_2 represent the data points that were low workload but were wrongly classified as high workload. The errors e_3 were the errors the majority voting scheme wrongly classified in either case. The overall performance results are displayed in Table 2. For two classifiers, the majority-voting scheme was considered inaccurate if both classifiers did not reach the same conclusion.

Table 2 – Performance of The Orthogonal Method versus Majority-Voting

Tests and Classifiers	e_1 errors	e_2 errors	e_3 errors
Test 1: C1 + C3	12	24	30
Test 2: C1 + C2	14	21	28
Test 3: C2 + C3	8	0	8
Test 4: C1+C2+C3	4	0	6

IV. Discussion of Results

From Table 2, some interesting results appear. When two classifiers are considered, the majority-voting scheme performs as well or better than the orthogonal method. As we go to higher dimensions, however, (Test 4), the combined effect of e_1 and e_2 errors is less for the orthogonal method as compared to the majority scheme. Also the Test 3 results are interesting because this is a poor estimator, yet the orthogonal projection scheme seems to include the relevant aspects of the decision-making space. The benefits of increasing the dimension of the orthogonal classifier seem to outweigh the benefits derived from the majority-voting scheme. As n gets larger, it appears this effect is more pronounced. Studies on ongoing to further investigate the dimensionality effect both within and across these candidate classifiers.

References

[1] A. Meystel, *Intelligent Systems – Annotated Bibliography and Survey*, Part 1-Systems with Intelligent Control, Part 2 – Systems with Learning, National Institute of Standards and Technology Report, 2000.

[2] D. W. Repperger, “Using Intelligent Machines to Modify or Adapt Human Behavior,” Chapter 9 in *Intelligent Machines: Myths and Realities*, CRC Press LLC, 2000.

[3] H-J Park, B. K. Kim, and K. Y. Lim, “Measuring the Machine Intelligence Quotient (MIQ) of Human-Machine Cooperative Systems,” *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, Vol. 31, No. 2, March, 2001, pp. 89-96.

[4] H. G. Kang and P. H. Seong, “Information Theoretic Approach to Man-Machine Interface Complexity Evaluation,” *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, Vol. 31, No. 3, May, 2001, pp. 163-171.

[5] D. W. Repperger and C. A. Russell, “A System to Improve the Quality of Information Gained from Multiple Data Sources,” *Proceedings of the IEEE 44th International Midwest Symposium on Circuits and Systems*, August 14-17, 2001, Dayton, Ohio.

[6] L. Lam and C. Y. Suen, “Application of Majority-Voting to Pattern Recognition: An Analysis of Its Behavior and Performance,” *IEEE Transactions on Systems, Man, and*

Cybernetics – Part A: Systems and Humans, Vol. 27, No. 5, September, 1997, pp. 553-568.

[7] T. Kailath, *Linear Systems*, Englewood Cliffs, New Jersey: Prentice-Hall, 1980.

[8] M. Mansuripur, *Introduction to Information Theory*, Prentice Hall, Inc., 1987.

[9] R. C. K. Lee, *Optimal Estimation, Identification, and Control*, The MIT Press, 1964.

[10] M. S. Grewal and A. P. Andrews, *Kalman Filtering - Theory and Practice*, Prentice Hall, 1993.

[11] T. B. Sheridan and W. R. Ferrell, *Man-Machine Systems – Information, Control, and Decision Models of Human Performance*, The MIT Press, 1974.

[12] Greene, K. A., Bayer, K. W., Kabrisky, M., Rogers, S. K., Russell, C. A., and Wilson, G. F., "A Preliminary Investigation of Selection of EEG and Psychophysical Features for Classifying Pilot Workload" *In Intelligent Engineering Systems through Artificial Neural Networks*, vol. 6. of *Proceedings of Artificial Neural Networks in Engineering International Conference*, 1996, pp. 691-697.

[13] E. Riess, "Automatic Monitoring of Electrical Parameters in the Semiconductor Industry Based on ROC," *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, Vol. 30, No. 6, November, 2000, pp. 853-857.

[14] J. Yen and R. Langari, *Fuzzy Logic – Intelligence, Control, and Information*, Prentice Hall, 1999.